"Meta Normalization for Text"

ABSTRACT OF THE DISCLOSURE

[0084] A system and method for normalizing encoded text data such as Unicode which is extensible without use of character definition tables through the use of metadata tagging. First, metadata characters, which have no effect on the interpretation of the raw text data, are used to express higher order protocols of encoded two text strings. Next, meta normal form conversion is performed on one or both of two strings to be compared, if both strings are not already in the same meta normal form. Finally, content equivalence determination is performed in which the characters in each string are compared to each other. If a string contains a metadata character, that character is ignored for purposes of equivalence comparison. The remaining characters represent the pure content of the string, e.g. characters without any particular glyph representation.